Improving Child-Adult Speaker Diarization During Non-Silent Segments

Richard Kha
University of Toronto

Bahar Aameri
University of Toronto

Abstract

During child-adult ASD therapy, metrics must be collected in order to evaluate the status and progress of the client. Collecting metrics manually takes attention away from providing care to the client, which may reduce the quality and effectiveness of the sessions for both client and therapist. Therefore, it would be better to automate some of this metric collecting process. A key component of language processing is speaker diarization; deciding "who said what" is important to calculate metrics like language complexity, response frequency, response rate, etc. Recent work in speech diarization in child-adult cases simplifies the diarization problem from deciding the particular speaker at a particular timestamp, to deciding if the particular speaker at a particular timestamp is a child or adult. A recent paper uses this property and introduces a model which improves on the diarization rate of the older models, decreasing diarization error in this specialized case. Our current work builds on this previous work by adjusting the output of the model in order to create higher accuracy child-adult diarizations during non-silent intervals. In particular, we improve the diarization accuracy of the recent model during non-silent intervals with a percent increase of 8% on the median of our test cases.

1 Introduction

Autism spectrum disorder (ASD) is a common neurological and developmental disorder which affects the behaviors of its carrier and the ability to socially interact with others [5]. The disorder is estimated to affect 0.76% of the world's population, with estimates of 1.68% of children aged 8 in the United States having a diagnosis [5].

Evaluation of ASD typically occurs during early childhood, where symptoms manifest as abnormal social behaviors compared to their peers [5]. After diagnosis, children with ASD usually receive behavioral treatment to relieve behavioral problems; in particular, in a study done in the United States, around 64% of children with autism received behavioral treatments [6]. Treatments commonly use applied behavior analysis (ABA) principles, as ABA is considered to be an evidence-based practice for children with ASD [4].

ABA involves improving behaviors based on seven core principles, which ensures that the treatment has properties such as being analytic and being effective [4]. The analytic requirement means that treatment must be reliable, thus, a collection of data to determine if behaviors are changing is necessary [3].

During child-adult ASD therapy, this data collection is usually done by the therapist. For example, during speech therapy, metrics like speech complexity can be used to gauge the progress and skill level of the client. However, as the therapist has to record these metrics manually, the therapist cannot devote their time solely to helping the client, which leads to a lower quality session for both the client and therapist. Therefore, it would be better to automate some of this metric collecting process.

When collecting language metrics, a key component of language processing is speaker diarization. Deciding "who said what" is important to calculate metrics like language complexity, response frequency, and response rate. Past audio diarization models are likely to have suboptimal accuracy on child-adult speech, as they were trained on mainly adult data. However, training diarization models on child-adult speech is difficult as child-adult datasets are scarce publicly because of concerns such as privacy.

A recent paper targeting child-adult speech diarization creates a new model which targets this very situation. Using a Whisper Encoder-based model, it converts the diarization problem to a speaker classification problem, where audio segments can only be classified as child or adult [8]. In their results, their model reduces diarization error rate over existing models when ran on their two private datasets.

In our work, we test the effectiveness of this diarization model on a public set of data (ASDBank AAC Minimally Speaking Autism Corpus), and provide a modification of this model which improves the accuracy of this diarization [7].

2 Custom diarization algorithm motivation

We compare Pyannote's speaker diarization model and the diarization model mentioned in the paper. Pyannote's speaker diarization model is a popular model used widely for speaker diarization. Anacdotally, the model seems to work poorly in the case of child-adult diarization, and takes a few minutes of processing time for an audio file a few minutes long. This is in contrast to the paper's model, which is built for child-adult diarization, and takes a few seconds of processing time for a similarly sized audio file.

Accuracy with silences: With the ASDBank AAC Minimally Speaking Autism Corpus, we first tested the accuracy of the diarization models on this dataset. We use the transcript files of the dataset as the ground truth, and compare it

, ,
.

, ,

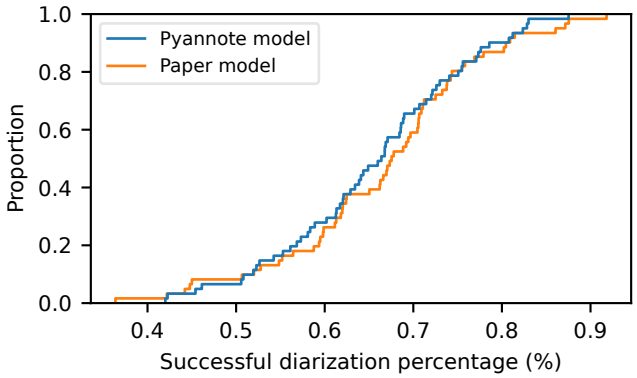

Figure 1. CDF of Diarization accuracy for the Pyannote model and Paper Model including silences

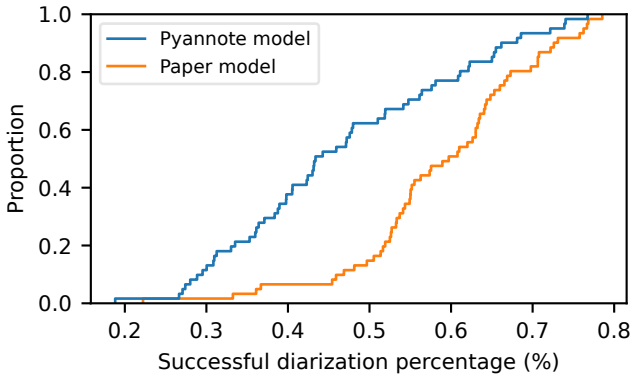

Figure 2. CDF of Diarization accuracy for the Pyannote model and Paper Model excluding silences

to the predicted diarizations of the models. The metric we use for this first test is accuracy, as in the total number of correct predictions over the total amount of diarizations. Specifically, we sum up the sub-intervals of correct predictions in the predicted diarization, and divide this by the sum of disjoint intervals in both the ground truth diarization intervals and the predicted diarization intervals. This accuracy accounts for false alarms, missed detections, and speaker confusion, so it is similar to diarization error, but is a measure of accuracy rather than error. Note that this accuracy measurement also accounts for silence detection accuracy for the models.

From results we see in figure 1, we see that the diarization accuracies of both models are similar. This is likely due to the silences in the audio files overpowering the accuracy, as often times, in audio files, there is more silence than speech.

Accuracy without silences: We therefore take the accuracy, without counting silences in the accuracy calculation. As well, we divide only by the sum of non-silent intervals in the ground truth diarization intervals. This measure of accuracy represents the accuracy of the model of determining the correct speaker when a speaker is actually speaking.

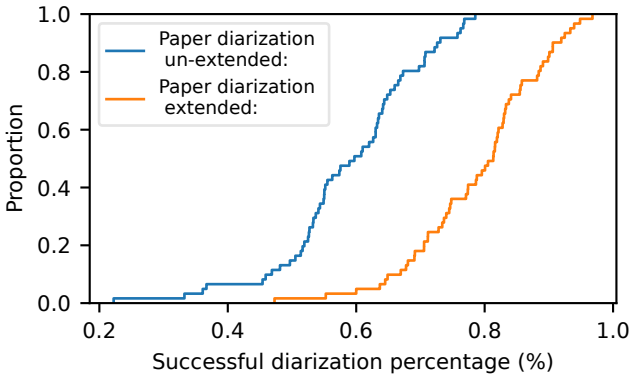

Figure 3. CDF of Diarization accuracy for the Paper model with and without the extension excluding silences

From results in figure 2, we see a large difference in accuracy between the pyannote model and the paper's model for a majority of the files. Specifically, we see a percent increase of 23% for the median of our test cases.

However, if we wanted to increase the accuracy of the model determining the correct speaker during non silence intervals, a simple way would be to make the prediction never predict silences, and predicts only child or adult. We propose the following rule:

- For every two sequential intervals in a diarization, extend both intervals to avoid the silence between them
- For example, if we have intervals [1,2,"child"] and [7,10, "adult], change the first interval to [1,4.5,"child"] and [4.5,10, "adult]

This creates new diarizations which have no silences within them. When testing accuracy during non-silence intervals, we get the graph in figure 3, which shows an increase in diarization accuracy over the old models. However, this makes the general accuracy (which includes silenced intervals) very low, as silences will no longer be predicted correctly.

Fortunately, because we did see an increase in diarization accuracy from extending intervals, this idea still has merit. If the intervals were extended, and combined with a method to classify silences, we can possibly get an increase in accuracy during non-silence intervals while maintaining the accuracy overall.

3 Custom diarization algorithm

The custom diarization algorithm has three main steps: getting the diarization intervals, getting the silence intervals, and intersecting the two sets of intervals to get a final diarization. The diarization algorithm works as follows:

- We first run the audio through the paper's model; this gives us a set of diarization intervals

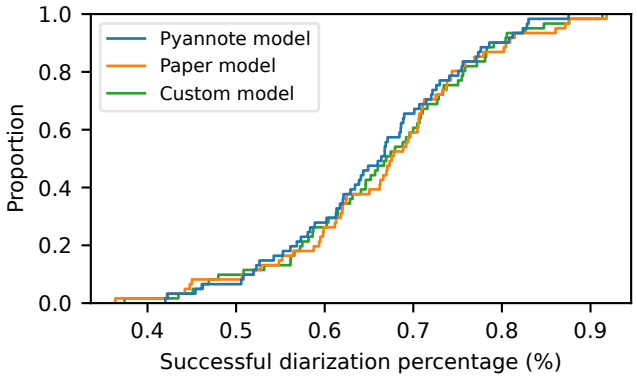

Figure 4. CDF of Diarization accuracy for all three models including silences

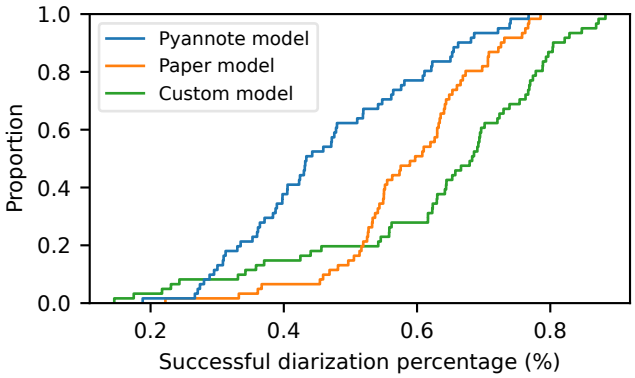

Figure 5. CDF of diarization accuracy for all three models excluding silences

- We extend these diarization intervals to get a set of diarization intervals without silences between the intervals
- We then run the audio through a deep neural network to remove the non-vocals
- We then threshold the audio and generate a set of diarization intervals which indicates when a person is speaking, and when is not
- We intersect the first set of diarization intervals (created from running the audio through the model) with the set of speaking intervals from the deep neural network to get a final set of intervals which represent the diarization

The deep neural network is necessary as thresholding would pick up non-vocals otherwise. Even with the deep neural network, the custom diarization algorithm runs at a similar speed to the paper's diarization algorithm, taking a few seconds to process an audio file of a few minutes length.

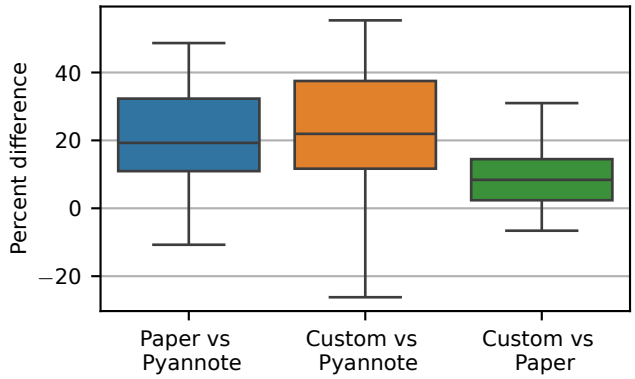

Figure 6. Boxplot of diarization accuracy for all three models excluding silences

4 Results and Discussion

Results:

The experiment was run on a personal computer. We test three models: the Pyannote model, the paper's model, and the custom diarization model. We ran three models on the data from the ASDBank AAC Minimally Speaking Autism Corpus [1].

In figure 4 we see the general diarization accuracy for all three models in the dataset. We see that the three models have roughly the same distribution of general diarization accuracy.

In figure 5 and 6 we see the non-silence diarization accuracy for all three models in the dataset. We see that the paper's model is more accurate than the pyannote model for a large portion of audio files, and we see that the custom diarization model is more accurate than the paper's model for a large portion of the audio files. Specifically, we see a percent increase of 8% for the median of test cases when comparing the custom diarization model to the paper's model.

Discussion: Our results show that our technique to improve the diarization accuracy of previous models by using detecting silences and extending intervals is promising for child-adult diarization. By extending the intervals, we gain accuracy when predicting speakers during non-silences, and combining this with voice detection techniques allows us to maintain our general accuracy which includes detecting silences.

Our extension of intervals is effective in this case as we are dealing with adult child diarization in particular. This makes our "guesses" for speakers accurate as there is a 50% chance we get it right while someone is speaking.

The deep neural network to remove non-vocals is usually criticized as it distorts the vocal audio as well. However, as we are simply using the audio to get intervals where vocals occur, this problem is not of concern with respect to our audio.

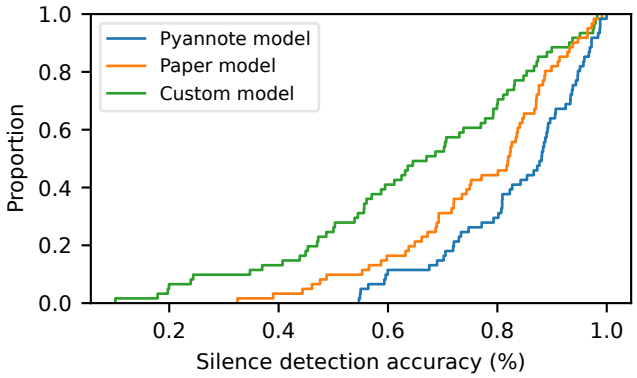

Figure 7. CDF of silence detection accuracy for all three models

We use medians and boxplot as they represent the accuracy change that happens to most of the files. They put less emphasis on the outliers.

Note that accuracy increased in the non-silent diarization accuracy case compared to the general diarization accuracy case. What is likely happening is that there is a trade off between accuracy during silences and accuracy during speaking occurring with this new algorithm. The new algorithm likely improves accuracy during non-silent intervals while decreasing accuracy during silent intervals. This can be further seen in figure 7 which shows that the custom model is worse at predicting silences. However, this trade off might be worth it in some cases, as accuracy during non-silent intervals is useful in areas such as transcription, where we need to bind certain words to certain speakers.

Implementation details: We use the DeepFilterNet noise suppressor [2]. When we feed the audio into a deep neural network to remove most non-vocal noises, it becomes a tensor. To compute the loudness signal at positions of this tensor, we use the root mean square of specific intervals around this position. We zero out portions of this tensor where there is an insignificant amount of noise with a threshold hyper-parameter. This hyper-parameter was chosen empirically.

To construct speech detection intervals, we move sequentially through the tensor. Intervals are continuous segments of the tensor where there is a significant amount of noise, and not too much of a gap between the significant noises. The size of the gaps we permit in an interval is also a hyper-parameter we chose empirically.

5 Future Work

From these preliminary results, we see that our custom model potentially improves on diarization accuracy during non-silent intervals. Our future work is to run this new model on other datasets to further validate our findings, and fine tune our models hyper-parameters for determining intervals when creating the vocal activity intervals. As well, we are looking into other voice activity detection methods in order to improve silence detection of the model, as this might lead to greater accuracy of the model in general. Finally, we are also looking into data synthesization techniques, such as the one mentioned in [8], to generate data which can be used for further fine tuning child-adult diarization models.

References

[1] [n. d.]. *ASDBank AAC Minimally Speaking Autism Corpus.* https://asd.talkbank.org/access/English/AAC.html

[2] [n. d.]. *DeepFilterNet.* https://github.com/Rikorose/DeepFilterNet

[3] Susan Palko Dawn Hendricks and Adam Dreyfus. [n. d.]. *What is Applied Behavior Analysis?* https://vcuautismcenter.org/resources/factsheets/printView.cfm/982

[4] H. Ho, A. Perry, and J. Koudys. 2021. A systematic review of behaviour analytic interventions for young children with intellectual disabilities. *Journal of Intellectual Disability Research* 65, 1 (2021), 11–31. https://doi.org/10.1111/jir.12780 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jir.12780

[5] Holly Hodges, Casey Fealko, and Neelkamal Soares. 2020. Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation. *Transl. Pediatr.* 9, Suppl 1 (Feb. 2020), S55–S65.

[6] Michael D Kogan, Catherine J Vladutiu, Laura A Schieve, Reem M Ghandour, Stephen J Blumberg, Benjamin Zablotsky, James M Perrin, Paul Shattuck, Karen A Kuhlthau, Robin L Harwood, and Michael C Lu. 2018. The prevalence of parent-reported autism spectrum disorder among US children. *Pediatrics* 142, 6 (Dec. 2018), e20174161.

[7] Aparna Nadig and Angela MacDonald-Prégent. 2024. *Augmentative and Alternative Communication Intervention Corpus: Minimally Speaking Autistic Children.* https://asd.talkbank.org/access/English/AAC.html

[8] Anfeng Xu, Tiantian Feng, Helen Tager-Flusberg, Catherine Lord, and Shrikanth Narayanan. 2024. Data Efficient Child-Adult Speaker Diarization with Simulated Conversations. arXiv:2409.08881 [eess.AS] https://arxiv.org/abs/2409.08881